DOCUMENT RESUME

ED 395 960 TM 025 098

AUTHOR Wingersky, Marilyn S.
TITLE A Consideration for Variable Length Adaptive Tests.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-40
PUB DATE Sep 89
NOTE 32p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Bayesian Statistics; Error of Measurement; *Estimation (Mathematics); Test Items; *Test Length; *True Scores
IDENTIFIERS *Asymptotic Standard Errors; Research Replication; *Stopping Rules

ABSTRACT

In a variable-length adaptive test with a stopping rule that relied on the asymptotic standard error of measurement of the examinee's estimated true score, M. S. Stocking (1987) discovered that it was sufficient to know the examinee's true score and the number of items administered to predict with some accuracy whether an examinee's true score was over- or underestimated. She theorized that this result might be due to the standard error being correct only asymptotically. J. B. Sympson (1985) recommended two Bayesian stopping rules that do not rely on asymptotic properties. This paper replicates the Stocking study using one of the variable-length adaptive testing procedures recommended by Sympson to see whether that procedure gives the same results as the Stocking study. The Sympson procedure uses a stopping rule that relies on the posterior standard deviation of the number-right true score on a criterion test. In both the Stocking study and Sympson procedure, knowing the examinee's true score and the number of items administered is sufficient for predicting whether the estimated true score is over- or underestimated. This is due to the fact that the magnitude of both the asymptotic standard error of measurement of the estimated scores and the posterior standard deviation of the true score varies as the estimated true score varies. (Contains two tables, three figures, and five references.) (Author/SLD)

RESEARCH REPORT

A CONSIDERATION FOR VARIABLE LENGTH ADAPTIVE TESTS

Marilyn S. Wingersky

Educational Testing Service
Princeton, New Jersey
September 1989

A Consideration for Variable Length Adaptive Tests*

Marilyn S. Wingersky

Educational Testing Service

Princeton, New Jersey

September 1989

---

*This study was supported by Educational Testing Service through Program Research Planning Council funding.

Abstract

In a variable-length adaptive test with a stopping rule that relied on the asymptotic standard error of measurement of the examinee's estimated true score, Stocking (1987) discovered that it was sufficient to know the examinee's true score and the number of items administered to predict with some accuracy whether an examinee's true score was over or under estimated. She theorized that this result might be due to the standard error being correct only asymptotically. Sympson (1985) recommended two Bayesian stopping rules that do not rely on any asymptotic properties. This paper replicates the Stocking study using one of the variable-length adaptive testing procedures recommended by Sympson to see whether that procedure gives the same result as found in the Stocking study. The Sympson procedure uses a stopping rule that relies on the posterior standard deviation of the number-right true score on a criterion test.

In both the Stocking study and the Sympson procedure, knowing the examinee's true score and the number of items administered is sufficient for predicting whether the estimated true score is over or under estimated. This is due to the fact that the magnitude of both the asymptotic standard error of measurement of the estimated scores and the posterior standard deviation of the true score varies

scores and the posterior standard deviation of the true score varies as the estimated true score varies.

A Consideration for Variable Length Adaptive Tests

Introduction

In computerized adaptive testing, individual examinees are presented with successive items at an appropriate level of difficulty depending upon each individual's preceding pattern of correct and incorrect responses. As a result, an individual's ability can be measured quite reliably with a relatively small number of items pitched at an appropriate level of difficulty for each individual. The number of items administered to an individual may be fixed or variable. If the number of items is variable, then some stopping rule must be used to determine when to stop administering items. Usually, the stopping rule states that items are administered until some variable, hereafter referred to as the stopping variable, becomes less than or greater than some cut-off value. Two different stopping rules are discussed in this paper; one, investigated by M. Stocking (1987), the other, developed by B. Sympson (1985).

Study 2 of Stocking (1987) investigated an adaptive testing procedure which stopped administering items when the estimated standard error of measurement of the estimated true score on a criterion test became less than some cut-off value. The standard

error of measurement computed by Stocking relies on the asymptotic information measure (Lord & Novick, 1968, p.457). Stocking found that if one knew an examinee's true score and the number of items administered to an examinee, one could predict whether the examinee's true score has been under or over estimated. This result is undesirable since low ability examinees with shorter tests could argue that they have been unfairly treated with respect to other low ability examinees with longer tests because shorter test lengths are associated with underestimation of test score. Similarly, high ability examinees with longer tests could challenge test fairness with respect to other high ability examinees with shorter tests because longer test lengths are associated with underestimation of test score (Stocking, 1987). She theorized that the result found in her study may be the consequence of the standard error being correct only asymptotically. Since very few items are given in an adaptive test, asymptotic properties are unlikely to be realized.

Sympson (1985) described a variable-length adaptive testing procedure that uses as the stopping variable either the posterior standard deviation of the number-right true score or the posterior standard deviation of the observed number-correct score on a criterion test. These standard deviations do not rely on asymptotic properties and can be computed accurately regardless of the number

of items administered. This paper investigates whether using the posterior standard deviation of the number-right true score as a stopping variable will give the same results as found in the Stocking study, namely, that knowing the number of items administered and the true score of an examinee are sufficient to predict whether an examinee's score has been over or under estimated. This paper is not meant to be a complete comparison of these two methods of variable length adaptive testing nor an indepth study of the Sympson procedure.

Methodology

The Model

The adaptive testing procedure used here is based on item response theory (IRT). IRT assumes that there is a mathematical function that relates the probability of a correct response on an item to an examinee's ability (Lord, 1980). The function, called an item response function, used in this paper is the three-parameter logistic model in which the probability of a correct response to item i given the ability θ, $P_i(\theta)$, is

$$P_i(\theta) = c_i + (1 - c_i)/(1 + \exp(-1.7a_i(\theta-b_i))), \qquad (1)$$

where a_i is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the discrimination power of the item, b_i is the point on the θ metric at the inflection point and represents the item difficulty, and c_i is the lower asymptote of $P_i(\theta)$.

Ability Estimates and Test Scores

In an adaptive test, the item parameters are known beforehand, and one wants to estimate the examinee's ability. Items are administered successively and the examinee's ability reestimated after each item to determine which item to administer next. The final ability estimate is based on the examinee's responses to the adaptive test, referred to by Sympson as the predictor test. While one may report estimates of θ that result from an adaptive test, such scores will have large errors of measurement for extreme (high or low) ability examinees. Instead of reporting θ, estimates of true scores on a specified criterion test can be reported. These estimated true scores have the desirable property that extreme true scores have small standard errors of measurement and also small posterior standard deviations. If an adaptive test is ended when the standard error of measurement or the posterior standard deviation of true scores falls below some cut-off value, the adaptive test will terminate quickly for examinees with extreme estimated true scores.

Stocking's Study 2

In Stocking's Study 2, ability was estimated from the adaptive test using maximum likelihood methods. However, the score used for reporting was the examinee's estimated true score on a criterion test computed using the ability estimated on the adaptive test. The criterion test used in this paper consists of all of the 120 items in her adaptive test item pool. For each examinee, this estimated true score, τ, was computed by the following formula

$$\hat{\tau} = \sum_{i=1}^{n} P_i(\hat{\theta}) \qquad (2)$$

where n is the number of items in the criterion test and $\hat{\theta}$ is the maximum likelihood estimate, MLE, of θ obtained from the adaptive test.

The stopping variable in Stocking's Study 2 was the estimated standard error of measurement of the number-correct true score on the criterion test computed using the ability estimate obtained from the adaptive test (Lord, 1983, equation 35). The adaptive test was terminated when the stopping variable became less than some cut-off value. This cut-off value was determined by the following procedure. A sample of approximately 5000 abilities was selected

from the abilities estimated for the examinees in a calibration of an administration of a real college placement test. These ability estimates (θ values) were converted to estimated number-right true scores (τ values) on the criterion test and the variance of these estimates computed. The cut-off for the standard error of measurement was chosen to be the square root of the error variance which, when used in conjunction with the variance of the estimated number-right true scores, would produce a reliability of .9 on the 120 item criterion test.

The Sympson Procedure

In the Sympson procedure, the estimated true score, τ^*, is the posterior mean of τ on the criterion test, given the response vector, V, on the adaptive, e.g. predictor, test, and is computed by evaluating

$$\tau^* = \mu(\tau|V) = [\int\{L(V|\theta)\ h(\theta)\}d\theta]^{-1}\ [\int\{L(V|\theta\ h(\theta)\tau(\theta)\}d\theta], \quad (3)$$

where

$h(\theta)$ is the prior density of θ.

$L(V|\theta)$ is the likelihood of observing response vector V on the items administered, given θ.

$\tau(\theta)$ is the number-correct true score on the criterion test and

is equal to $\sum_{i=1}^{n} P_i(\theta)$.

The posterior variance of τ on the criterion test, given the observed vector of item responses on the predictor test, is the stopping variable. The formula for the posterior variance is

$$\sigma^2(\tau|V) = \mu(\tau^2|V) - [\mu(\tau|V)]^2 \tag{4}$$

where

$$\mu(\tau^2|V) = [\int(L(V|\theta)h(\theta))d\theta]^{-1} [\int(L(V|\theta)h(\theta)[\tau(\theta)]^2)d\theta] \tag{5}$$

Sympson's formula for the posterior variance does not rely on any asymptotic properties and can be computed at any test length as long as the numerical quadrature procedure that is used to evaluate Equations 3 and 4 is accurate.

The adaptive test is stopped when the posterior variance becomes less than some cut-off value, say σ_c. This cut-off value is determined such that, if the regression between the estimated true scores and the actual true scores is linear, the squared linear correlation between the two sets of scores in the prior population

will be greater than some specified value. The reasoning behind the computation of the cut-off value follows.

The squared eta coefficient for estimating τ from V is the same as the squared linear correlation between the estimated true scores and the actual true scores providing the regression between the two sets of scores is linear. The formula for η^2 is given by

$$\eta^2(V \rightarrow \tau) = 1 - \frac{E[\sigma^2(\tau|V)]}{\sigma^2(\tau)} \tag{6}$$

where

$E[\sigma^2(\tau|V)]$ is the expectation of $\sigma^2(\tau|V)$ over all possible V.

$$\sigma^2(\tau) = \mu(\tau^2) - [\mu(\tau)]^2 \tag{7}$$

$$\mu(\tau) = \int\{h(\theta)\tau(\theta)\}d\theta \tag{8}$$

$$\mu(\tau^2) = \int\{h(\theta)[\tau(\theta)]^2\}d\theta \ . \tag{9}$$

Since $\sigma^2(\tau)$ is fixed for a given population, imposing an upper bound, the cut-off value σ_c, on $\sigma^2(\tau|V)$ for all possible V puts an upper bound on $E[\sigma^2(\tau|V)]$. This upper bound has the effect of putting a lower bound on η^2. σ_c is computed by substituting σ_c^2 for $E[\sigma^2(\tau|V)]$ in equation 6, setting η^2 to the desired lower bound and solving for σ_c.

$$\sigma_c = [(1 - \eta^2)\sigma^2(\tau)]^{1/2} .$$

For this study η^2 was set to .90.

Computer Program

The computer program used in this study, (SIMCAT2) was originally written by Sympson and modified by the author to handle a larger criterion test, to select items using the item selection algorithm described in the next section, to use 48 point Gauss quadrature and to execute more quickly. The Bayesian modal estimate of ability, assuming a normal prior distribution with mean zero and standard deviation one, was used in the adaptive test item selection algorithm described in the next section. In this study, $h(\theta)$ was specified to be a normal distribution with mean zero and standard deviation one.

Data

The CAT Item Pool and Item Selection Algorithm

The item parameters used in this study were the same as those used in Stocking (1987, Study 2) which were originally obtained in a calibration of items in the Reading Comprehension subtest of the New Jersey College Basic Skills Placement Testing program. There were 120 items in the adaptive test item pool. This item pool contained item difficulty levels ranging from -1.04 to 1.27. The criterion test was defined to contain all 120 items in the item pool. Thus $\hat{\tau}$ and τ^* were estimates of an examinee's true score on the entire item pool, obtained from abilities estimated from the responses to a subset of the items in the pool.

The item selection algorithm was as follows, the first item to be administered was selected randomly from a group of five items that have the maxima of their information functions (see Stocking, 1987, eq. 3) at a middle ability level around zero. The second item was selected randomly from a group of three items which are maximally informative at an extremely low (high) ability level if the first item was answered incorrectly (correctly). This method of selecting the second item is necessary to obtain a maximum likelihood ability estimate which requires at least one correct and one incorrect response. This method was not necessary for the

Sympson procedure, but was used to keep the results from the two procedures as similar as possible. The remaining items were selected to have maximal information at the current estimate of ability. Since the original test contained several different types of items, care was taken to balance the administration of the different item types across ability levels as was done in the Stocking study.

As in the Stocking study, the Sympson procedure was run on 1800 simulated examinees, 200 examinees at each of 9 true score points on the 120 item criterion test. These scores ranged from 30 to 110 in increments of 10. This range corresponds to θ from -2.46 to 1.53.

As in Stocking Study 2, the minimum number of items was set to 10 for the adaptive test. This was required since, for extreme examinees, the standard errors were sufficiently small that if no minimum were set, as few as 4 items might be given which would be insufficient to obtain a reasonable MLE estimate of ability or to administer a sufficient mix of item types. The maximum number of items was set at 40 as being a sufficiently long test.

Results

The first part of this section checks that the cut-off criterion used in Sympson's Bayesian procedure actually gives the squared correlation that one wants. The second part looks at the

conditional bias and the precision of the two procedures. The third part discusses the answer to the question, "Using the Sympson procedure, can one predict whether an examinee's ability is over or under estimated if one knows the true score and the number of items administered?"

To check that the Sympson procedure, as implemented in the modified version of SIMCAT2 actually gave a squared fidelity coefficient of .90, a random sample of 1000 simulated examinees was selected from a standard normal population and run on the procedure stopping the adaptive test when $\sigma(\tau|V)$ became less than the σ_c that would give an η^2 greater than .90. For this sample, the squared linear correlation between τ^* and τ was .906.

To be sure that the answer to the main question addressed in this paper using Sympson's procedure is comparable to the answer found in Stocking Study 2, it is necessary to make sure that Sympson's procedure recovers the true values approximately as well as they were recovered in Stocking's Study 2. For this pupose, the conditional bias and precision of the estimated scores given the true scores will be used. It should be remembered that the Sympson Bayesian procedure is designed to be optimal over the distribution of abilities used in the prior distribution, and will not be optimal when looking at conditional distributions.

In this paper the precision of the estimates used is the root mean square error, RMSE, between the estimated true score and the actual true score averaged over all examinees at a given true score. The RMSE is also broken down into its two components, the bias and the standard error of the estimates. The formulas for the RMSE, bias, and the standard error of measurement, (SEM), are

$$\text{RMSE(Stocking)} = \left[\frac{1}{N} \sum_a (\hat{\tau}_a - \tau)^2 \right]^{1/2} \tag{10}$$

$$\text{Bias(Stocking)} = \frac{1}{N} \Sigma (\hat{\tau}_a - \tau)$$

$$\text{SEM(Stocking)} = \left[\frac{1}{N} \Sigma \{ (\hat{\tau}_a - \tau) - \text{Bias} \}^2 \right]^{1/2}$$

$$\text{RMSE(Sympson)} = \left[\frac{1}{N} \sum_a (\tau^*_a - \tau)^2 \right]^{1/2} \tag{11}$$

$$\text{Bias(Sympson)} = \frac{1}{N} \Sigma (\tau^*_a - \tau)$$

$$\text{SEM(Sympson)} = \left[\frac{1}{N} \Sigma \{ (\tau^*_a - \tau) - \text{Bias} \}^2 \right]^{1/2} .$$

The summation is over the examinees at a given true score, τ. N is the number of examinees with this true score.

Table 1 contains these statistics for the different true scores for the variable-length test using Sympson's procedure and for the variable length-test in Stocking's Study 2. For most of the scores,

both procedures do approximately as well in terms of RMSE. For the two extreme score groups, Sympson's procedure gave a higher RMSE than was observed in Stocking's Study 2 procedure. This result is an effect of Sympson's Bayesian method of estimating τ, which regresses extreme estimates towards the prior mean of τ. This increases the bias component in the conditional mean square error while reducing the overall mean square error in the population defined by $h(\theta)$. As expected, the bias is in the opposite direction for the two procedures. Table 2 contains the average number of items for each of the true scores for the two procedures. It is interesting to note that Sympson's procedure required fewer items to produce estimated scores that were as good as the estimated scores in Stocking study 2 for most of the scores.

Insert Table 1 and Table 2 about here

Given that the estimated scores using Sympson's procedure are reasonable, the question "Using Sympson's procedure, can one predict whether an examinee's ability is over or under estimated when one knows the true score and the number of items administered?" can now be answered. Figures 1 and 2 graphically display the answer. Figure 1 contains frequency distributions of the number of items

administered for the underestimated abilities, (u), and for the overestimated abilities, (o), at each score level. The median is the middle line in each box, the bottom and top of the box are the 25th and the 75th percentiles, respectively. The end of the line below the box is the 10th percentile and the end of the line above the box is the 90th percentile. Some of the boxes are bounded by the lower bound of 10, the minimum number of items administered. The top plot contains the distributions from Stocking Study 2. The bottom plot contains the distributions from the Sympson Bayesian procedure. The distributions for the two extreme scores of 30 and 110 have been omitted because they reflected floor and ceiling effects. For both methods, except for the middle score, if one knows an examinee's true score and the number of items administered, one can predict with a fair degree of certainty whether the estimated score is too high or too low. The prediction is a little less certain for the Sympson Bayesian procedure than for the Stocking Study 2 procedure. This is perhaps seen better by the plots in Figure 2. The top plot is for the Stocking Study 2 procedure, the bottom plot is for the Sympson procedure. Here, the number of items administered is plotted against the residual for all of the examinees with a true score of 50. The residual is the estimated true score minus the true score. For both procedures, the plots show that, as the

residual increases from negative to positive, the number of items administered increases, although the slope is not quite as steep for the Sympson procedure as for the procedure in Stocking Study 2.

Insert Figures 1 and 2 about here

The correlation between errors of estimate and the test length that was observed in Stocking's Study 2 is, therefore, not due to the failure of the asymptotic estimate of the standard error of measurement but is a problem with the stopping criterion. In both Stocking's Study 2 and Sympson's procedures the variable that is used to determine when to stop administering items drops rapidly as the estimate of the true score approaches its bounds. In either tail of the true score range, if the estimate of the true score is nearer the limit of the range, the stopping variable will be smaller and, therefore, the adaptive test will stop sooner than if the estimated true score were nearer the middle of the range. Changing to a different metric, such as the θ metric, for reporting scores will not remove this problem as long as the magnitude of the stopping variable is related to the reported score.

An adaptive testing procedure may be viewed as a sequential estimation algorithm. In typical sequential estimation problems,

the parameter to be estimated is usually transformed to have a constant variance if the sample (here the test length) is large enough. However, this is not possible in adaptive testing, since a different transformation would be required for each examinee and the scores would no longer be comparable (Stocking, 1987).

The ability to predict whether the true score is over or under estimated from knowing the true score and the number of items administered is a drawback of both of these variable-length adaptive testing procedures.

However, it is questionable whether this problem is a serious drawback, since, in practice, only the examinee's estimated score is known and not the true score. If only the examinee's estimated score and the number of items administered are known, can one predict whether the estimated score is below or above the true score? For the rectangular distribution of abilities used, the answer is no. Whether this answer generalizes to other distributions of ability was not investigated. Figure 3 contains the distributions of the number of items administered plotted against the estimated score, split into overestimated scores and underestimated scores. The top plot is for the Stocking Study 2 procedure; the bottom plot is for the Sympson procedure. For each box, the median test length is the line across the middle of the

box, the top of the box is the 75th-percentile test length, the bottom of the box is the 25th-percentile test length. The end of the line extending below the bottom of the box is the 10th-percentile test length. The end of the line extending above the top of the box is the 90th-percentile test length. To produce plots similar to the true score plots, only estimated scores that were multiples of 10 were plotted. The boxes at each score included the estimated scores that were "close" to the plotted score. "Close" was defined as ± 3 from the plotted score. Thus, for example, the scores included in the two boxes plotted near 70 ranged from 67 up to, but not including, 73. The box plotted with a "u" at the bottom represents the interquartile range of test lengths among examinees who received an estimated score that was less than the examinee's true score. The box plotted with an "o" at the bottom represents the interquartile range of test lengths among examinees who received an estimated score that was greater than the examinee's true score. The number of examinees represented in each group varies from 50 to 148. For the Stocking Study 2 data, the number of examinees ranged from 36 to 72. These plots show that the number of items administered and the estimated score are not sufficient to predict whether the estimated score is above or below the examinee's true score. For example, for low scores, the median of the

underestimated scores is lower than the median of the overestimated scores. However, the median number of items for the under (over) estimated scores is within the interquartile range of the over (under) estimated score.

Insert Figure 3 about here

Conclusion

In this project, the analysis done in Study 2, Stocking (1987) was repeated with a stopping rule that uses the posterior standard deviation of the number-right true score on a criterion test as suggested by Sympson (1985). This standard deviation can be computed accurately and does not rely on any asymptotic properties. Using this stopping variable did not remove the problem with variable length adaptive testing found in the Stocking study, that is, if one knows the true score of the examinee and the number of items administered, one can predict with some accuracy whether an examinee's score was over or under estimated. This problem is caused by using a stopping variable that varies as the estimated score varies. It is questionable whether this problem is a serious drawback to variable length adaptive testing since all that is known in practice is an examinee's estimated score and not his true score.

Knowing only the estimated score and the length of the adaptive test, one can not predict with any certainty whether the true score has been over or under estimated.

References

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley Publishing Company.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-245.

Stocking, M. S. (1987). Two simulated feasibility studies in computerised adaptive testing. Applied Psychology: An International Review, 36, 263-277.

Sympson, J. B. (1985). Bayesian estimation of true scores and observed scores on a criterion test. Paper presented at the annual meeting of the Psychometric Society, Nashville, TN.

Table 1

Conditional Bias and Precision of

Sympson's Bayesian Procedure and the Stocking Study 2 Procedure

	Root Mean Square Error		Bias		Standard Error of Measurement	
ξ	Sympson	Stocking	Sympson	Stocking	Sympson	Stocking
110	5.8	4.8	-4.2	0.0	4.0	4.8
100	5.9	5.8	-1.3	1.6	5.8	5.6
90	6.8	7.5	-1.1	1.9	6.7	7.3
80	7.8	7.6	0.7	1.8	7.8	7.4
70	7.5	7.8	-0.7	1.4	7.5	7.7
60	7.9	7.9	0.7	-0.4	7.9	7.9
50	6.9	7.5	1.6	-0.4	6.8	7.5
40	6.2	6.0	3.4	-1.0	5.2	5.9
30	8.7	3.9	7.9	0.9	3.8	3.8

Table 2

Average Number of Items for

Sympson's Bayesian Procedure and the Stocking Study 2 Procedure

	Average Number of Items	
ξ	Sympson	Stocking
110	10.7	10.8
100	13.4	14.2
90	17.4	20.5
80	20.6	25.2
70	22.5	28.3
60	21.7	27.9
50	18.3	23.3
40	13.7	16.6
30	11.6	13.4

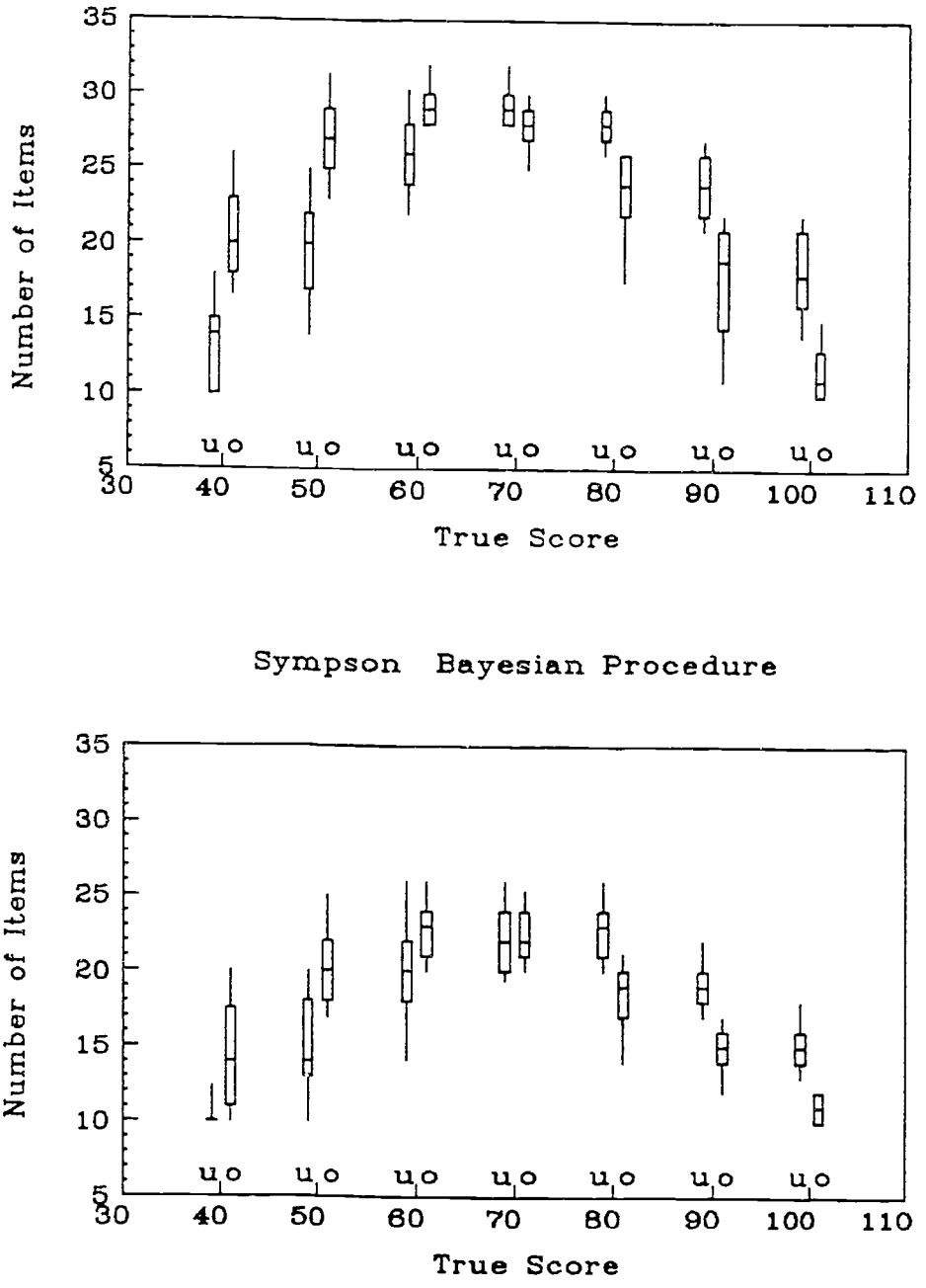


Figure 1. The distributions of the number of items administered for the underestimated and overestimated scores for each of the selected true scores for the two procedures. The distributions for the underestimated scores have a "u" below the box, the distributions for the overestimated scores has an "o" below the box. The distributions are graphed using the following percentiles:

10th percentile - the end of the line below the box
25th percentile - the bottom of the box
50th percentile - the line inside the box
75th percentile - the top of the box
90th percentile - the end of the line above the box

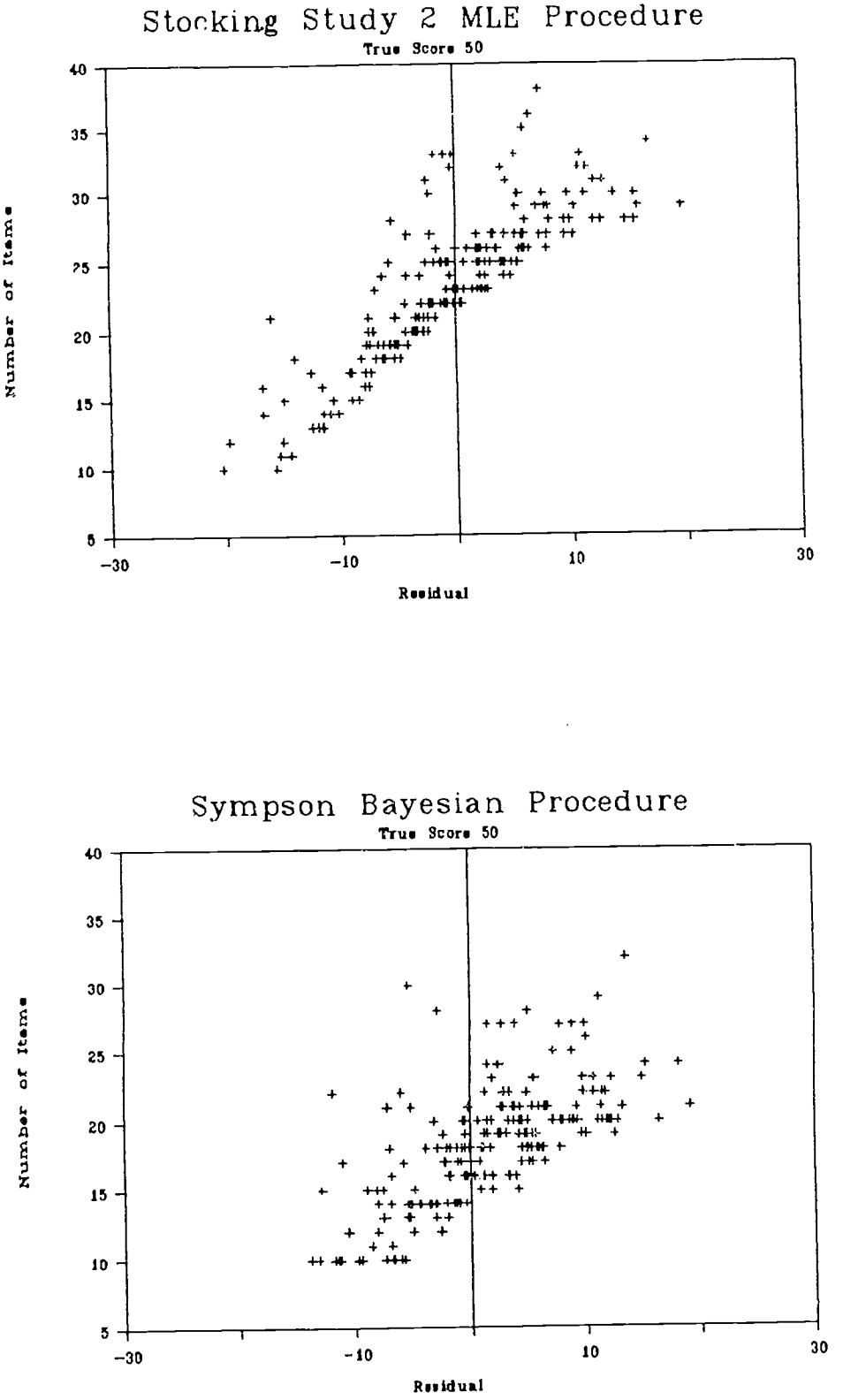
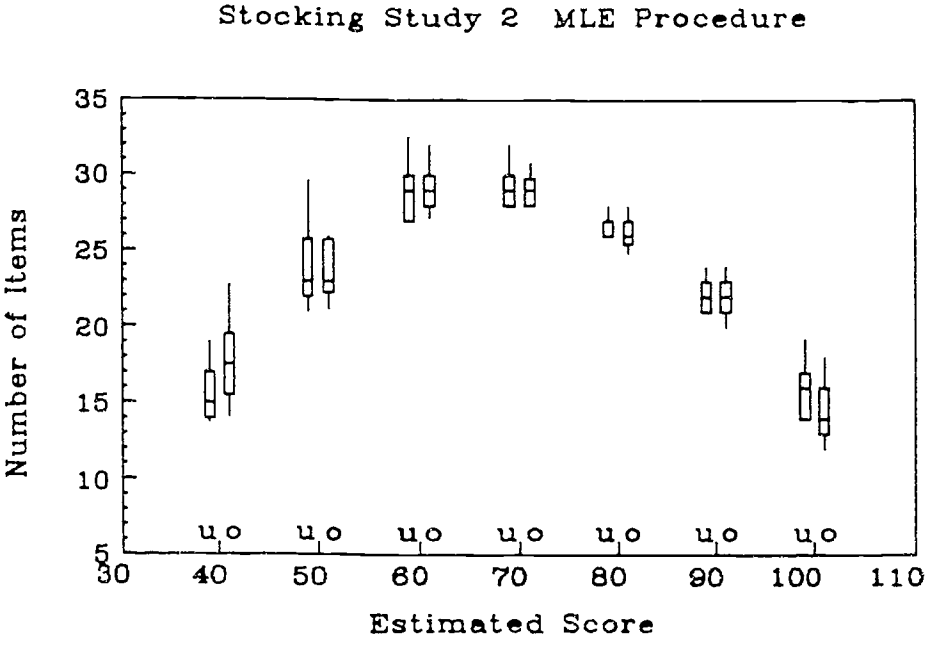


Figure 2. The number of items administered plotted against the residuals for the set of examinees with a true score of 50. In the top plot for the Stocking Study 2 procedure, the residual is $\hat{\tau} - \tau$. In the bottom plot for the Sympson procedure, the residual is $\tau^* - \tau$.



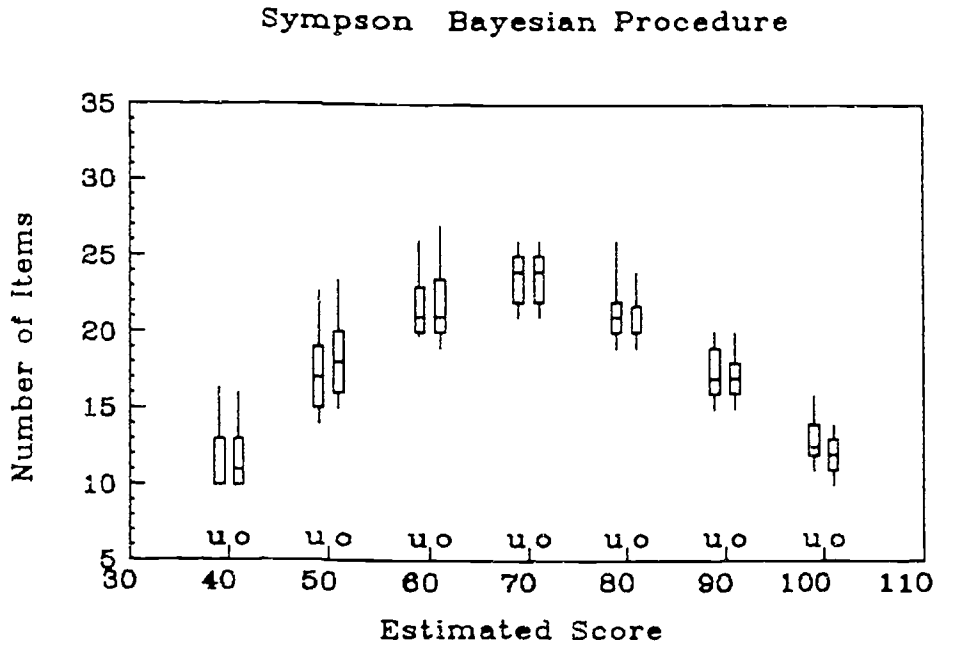


Figure 3. Frequency distributions of the number of items administered for the underestimated and overestimated true scores. The examinees included at a particular score are the ones whose esitmated score was within a range of ±3 of the score. The distributions for the underestimated scores have a "u" below the box; distributions for the overestimated scores have an "o" below the box. The distributions are graphed using the following percentiles:

10th percentile - the end of the line below the box
25th percentile - the bottom of the box
50th percentile - the line inside the box
75th percentile - the top of the box
90th percentile - the end of the line above the box